Journal of Computer Applications
计算机应用,2021,41(7):1922 - 1927
ISSN 1001-9081
CODEN JYIIDU
2021-07-10
http://www.joca.cn

文章编号:1001-9081(2021)07-1922-06
DOI:10.11772/j.issn.1001-9081.2020081214

基于深度对齐网络的生成对抗网络伪造人脸检测

汤桂花*,孙　磊,毛秀青,戴乐育,胡永进
(信息工程大学,郑州 450001)
(*通信作者电子邮箱tgh1678348898@stu.xjtu.edu.cn)

摘　要:针对现有的生成对抗网络(GAN)伪造人脸图像检测方法在有角度及遮挡情况下存在的真实人脸误判问题,提出了一种基于深度对齐网络(DAN)的GAN伪造人脸图像检测方法。首先,基于DAN设计面部关键点提取网络,以提取真伪人脸关键点位置;然后,采用主成分分析(PCA)方法将每一组关键点映射到三维空间,从而减少冗余信息以及降低特征维度;最后,利用支持向量机(SVM)五折交叉验证对特征进行分类,并计算准确率。实验结果表明,该方法通过提高面部关键点定位准确度改善了由于定位误差引起的面部不协调问题,进而降低了真实人脸误判率。与VGG19、XceptionNet和Dlib-SVM方法相比,正脸情况下,该方法的ROC下面积(AUC)值提高了4.48到32.96个百分点,平均精度(AP)提高了4.26到33.12个百分点;有角度及遮挡人脸情况下,该方法的AUC值提高了10.56到30.75个百分点,AP提高了7.42到42.45个百分点。

关键词:生成对抗网络;深度对齐网络;面部关键点;图像伪造;伪造检测
中图分类号:TP391.41　　**文献标志码**:A

Generative adversarial network synthesized face detection based on deep alignment network

TANG Guihua*, SUN Lei, MAO Xiuqing, DAI Leyu, HU Yongjin
(*Information Engineering University*, *Zhengzhou Hernan* 450001, *China*)

Abstract: The existing Generative Adversarial Network (GAN) synthesized face detection method has misjudgment of real faces with angles and occlusion, therefor a GAN-synthesized face detection method based on Deep Alignment Network (DAN) was proposed. Firstly, a facial landmark extraction network was designed based on DAN to extract the locations of facial landmarks of genuinus and synthesized faces. Then, in order to reduce the redundant information and feature dimensionality, each group of landmarks was mapped to the three-dimensional space by using the Principal Component Analysis (PCA) method. Finally, the features were classified by using 5-fold cross-validation of Support Vector Machine and the accuracy was calculated. Experimental results show that the proposed method improves the face dissonance caused by location errors by improving the accuracy of facial landmark location, which reduces the misjudgment rate of real faces. Compared with VGG19, XceptionNet and Dlib-SVM methods, this proposed method has the Area Under Receiver Operating Characteristic curve (AUC) increased by 4.48 to 32.96 percentage points and Average Precision (AP) increased by 4.26 to 33.12 percentage points on frontal faces; and has the AUC increased by 10.56 to 30.75 percentage points and AP increased by 7.42 to 42.45 percentage points on faces with angles and occlusion.

Key words: Generative Adversarial Network (GAN); Deep Alignment Network (DAN); facial landmark; image synthesis; forgery detection

0　引言

传统图像伪造技术通常利用图像编辑软件(如Adobe Photoshop、Illustrate和GIMP等)通过复制粘贴,对真实图像进行拼接,有针对性地改变图像内容,伪造图像。随着人工智能、云计算、大数据技术的快速发展,以及高性能计算硬件的不断更新迭代,通过训练大量数据伪造出高质量图像的深度伪造技术成为主流。Goodfellow等[1]于2014年提出的生成对抗网络(Generative Adversarial Network, GAN)提供了一种基于零和博弈的高质量图像伪造方法,能够直接由随机噪声通过深度神经网络生成图像。原始GAN经过不断改进发展,已经可以生成高分辨率的逼真图像。PGGAN(Progressive Growing of GAN)[2]是伪造高分辨率人脸图像的重大突破,通过渐进增长的训练方式,可以在CelebA-hq数据集的训练下生成1 024×1 024的高分辨率人脸图像,是目前伪造图像质量最高的模型之一。STGAN(Style-Transfer GAN)[3-4]在PGGAN模型的基础上作了进一步改进。然而,STGAN解决的是人脸之间的风格转换问题,在已有人脸图像的基础上进行伪造,并非

收稿日期:2020-08-12;**修回日期**:2020-10-23;**录用日期**:2020-11-13。　**基金项目**:国家重点研发计划项目(2016YFB0501900)。
作者简介:汤桂花(1996—),女,四川乐山人,硕士研究生,主要研究方向:图像处理、机器视觉;　孙磊(1973—),男,江苏靖江人,教授,博士,主要研究方向:网络空间安全、可信计算;　毛秀青(1980—),男,安徽滁州人,副教授,硕士,主要研究方向:信息安全;　戴乐育(1990—),男,河南郑州人,讲师,硕士,主要研究方向:神经网络加密、硬件加速;　胡永进(1981—),男,山东潍坊人,讲师,硕士,主要研究方向:主动防御、态势感知。

直接由随机噪声生成图像,与以往GAN图像伪造模型有着本质的不同,因此本文选择PGGAN伪造人脸图像作为评估数据集,能更好地代表目前GAN伪造技术的水平和揭示GAN伪造图像的本质差异。

作为一项新颖的科技创新,深度伪造技术在带来全新产业价值的同时也给个人和社会带来了相应的风险和挑战。近年来,日益开放的网络环境为伪造信息的传播创造了理想空间,英、法、美等国相继出现了利用深度伪造技术制造假新闻诱导舆论、欺骗公众甚至进行间谍活动的事件,引起人们对伦理、法律和安全方面的重大担忧。传统的伪造图像检测技术通常对图像篡改操作中所引入的特有效应进行分析,并定位出篡改区域。GAN伪造图像与传统伪造图像相比,不存在图像拼接以后留下的痕迹,所以传统的伪造图像检测技术无法有效甄别GAN伪造图像。目前针对GAN伪造图像也提出了一些有效的检测方法。Li等[5]通过分析H、S、V和Cb、Cr、Y通道中真伪图像颜色分量的差异,提出了一种彩色图像统计特征集来检测GAN伪造图像;McCloskey等[6]对一种常用GAN模型中的生成网络结构进行分析发现,GAN伪造图像的饱和像素频率和颜色分量统计关系与相机拍摄的图像不同,利用这两个线索作者设计了针对GAN伪造图像的检测算法;Nataraj等[7]则通过提取像素域中三个颜色通道上的共现矩阵表示颜色分量间的相关性,并结合深度学习的方法来检测GAN伪造图像。然而面对图像合成后的后处理,这些检测方法的鲁棒性较差,如常见的压缩、添加噪声、模糊等会使图像的像素值产生变化,像素间的相关性会遭到破坏,GAN伪造图像和真实图像间的上述差异就会被大大降低,影响检测效果。Tariq等[8]提出了基于深度卷积神经网络(Convolutional Neural Network, CNN)的分类器来检测GAN伪造的人脸图像,虽然也取得了良好的分类效果,但未从GAN伪造图像本身的特性进行分析,训练好的模型面对其他GAN伪造图像的分类效果不稳定,并且需要大量数据和较高的硬件条件;Yang等[9]通过对一定量的GAN伪造人脸图像观察发现,基于GAN的人脸伪造算法可以生成具有高度真实细节的人脸器官,但对人脸中这些器官的位置却没有明确的约束,导致即使是使用先进的PGGAN生成的人脸图像也存在面部不协调的问题,如图1所示。可以观察到图1的前两张人脸轮廓均存在不同程度的扭曲,第三张人脸的右眼呈现不自然的闭合。该文进一步利用Dlib函数库中集成的人脸检测器提取了68个面部关键点位置,并将位置数据作为特征向量对真伪人脸进行分类,在正脸图像中取得了较好的效果;但是在有角度及遮挡情况,由于定位偏差人为引入了面部不协调,真实人脸就容易被误判为GAN伪造的不协调伪造人脸,导致判真率较低。

图1　PGGAN生成的不协调人脸示例

Fig. 1　Examples of dissonant faces synthesized by PGGAN

针对真实人脸的误判问题,本文提出了一种基于深度对齐网络(Deep Alignment Network, DAN)[10]的GAN伪造人脸图像检测方法。基于DAN构建的面部关键点提取网络通过输入人脸的全局信息避免进行关键点定位时被局部信息误导,提升了面部关键点定位准确度,能有效改善在有角度及遮挡情况下真实人脸误判为伪造人脸的问题。本文方法在PGGAN伪造的高分辨率人脸数据集和CelebA(CelebFaces Attributes Dataset)[11]人脸数据集上进行了实验,结果表明该方法能够有效提高有角度及遮挡情况下真实人脸的检测正确率。

1　基于DAN的GAN伪造人脸图像检测方法

1.1　本文检测方法框架

本文在复现文献[9]的工作并对误判的真实人脸图像进行关键点定位后分析发现,误判人脸的关键点定位均出现了较大偏差。图2为误判人脸的定位情况,在有角度和遮挡情况下,面部器官及轮廓定位出现较大偏差,人为引入了面部不对称、不协调,形成如图1所示的PGGAN不协调人脸,从而导致真实人脸被判定为GAN伪造的不协调人脸,产生真实人脸误判情况。Dlib提取68个面部关键点是基于ERT(Ensemble of Regression Trees)算法[12],即基于梯度提高学习的回归树方法。该算法以标准人脸关键点作为初始值,每次从关键点附近进行特征采样取值,通过迭代来修正关键点的位置,并且各特征点是独立的回归,没有利用相互之间的相对位置关系。而这种仅基于局部区域进行关键点定位的方法,缺少整体信息,在面部存在遮挡和旋转角度的情况下容易被局部信息误导,导致定位出现偏差。这种由于定位偏差引入的面部不协调进一步造成在真伪人脸分类中真实人脸的误判问题。

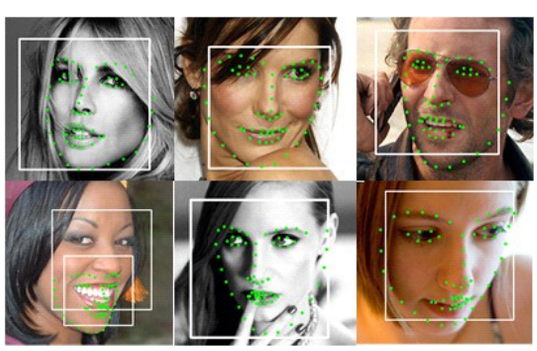


图2　误判人脸的关键点定位情况

Fig. 2　Location of landmarks of misjudged faces

因此针对文献[9]中提出的GAN伪造人脸检测方法在真实人脸部分遮挡及有角度情况下存在误判的问题,本文方法通过提高面部关键点的定位准确度,降低遮挡及有角度情况下的面部关键点定位偏差,改善真实人脸误判问题,提升判真率。

本文选择68个面部关键点对真伪人脸进行定位,利用关键点位置分类暴露GAN伪造人脸图像存在的面部不协调问题。68个关键点分为内部关键点和轮廓关键点,内部关键点包含眉毛、眼睛、鼻子、嘴巴等共计51个关键点,轮廓关键点包含17个关键点,可以较为精确地定位面部器官及轮廓。检测方法的整体流程如图3所示,所用人脸图像均来自于公开数据集。

首先基于DAN构建面部关键点提取网络,并利用标准人脸对齐数据集对其进行训练,保存模型。然后利用面部关键点提取模型对人脸图像进行定位并输出关键点位置信息:为了减少冗余信息的同时最大化保持数据的内在信息,选择主成分分析(Principal Component Analysis, PCA)[13]降维方法将136维的真伪人脸图像面部关键点特征映射到三维空间,以便进行下一步分类。最后,利用支持向量机(Support Vector

Machine，SVM）构建一个分类器以区分GAN伪造的和真实的人脸，将处理后的数据作为特征向量输入分类系统进行训练测试并输出分类结果。

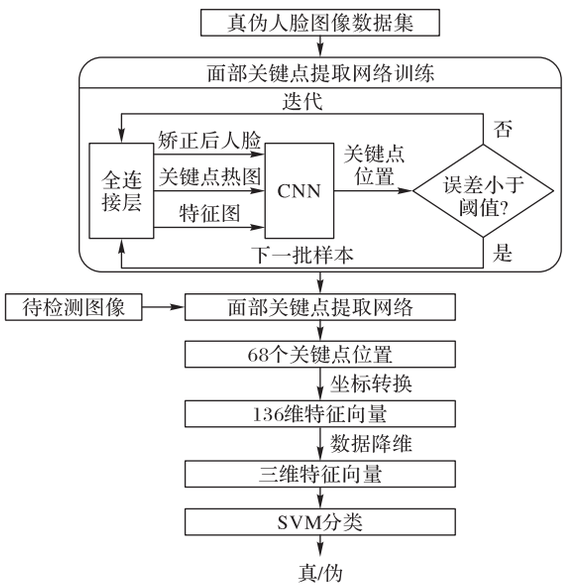


图3 GAN伪造人脸图像检测方法整体流程

Fig. 3 Flowchart of GAN-synthesized face detection method

1.2 面部关键点提取网络

本文基于DAN构建真伪人脸图像的面部关键点提取网络，其中前馈神经网络架构如表1所示，Conv表示卷积层。在网络的第三、四层，受Inception-v3[14]的启发，将原网络的3×3卷积分解为1×3和3×1卷积，提高网络非线性的同时降低网络的参数和计算量；采用全局均值池化（Global average pooling，Avgpool）代替全连接层（Fully Connected layer，FC），减少网络参数，避免过拟合。除了最大池化层和输出层之外，每一层都进行批处理规范化，并使用线性整流函数（Rectified Linear Unit，ReLU）进行激活，最后一层输出当前的关键点位置估计。

表1 前馈神经网络架构

Tab. 1 Structure of feed-forward neural network

类型	输入尺寸	输出尺寸	卷积核
Conv1a	112×112×1	112×112×64	3×3,1
Conv1b	112×112×64	112×112×64	3×3,1
Pool1	112×112×64	56×56×64	2×2,2
Conv2a	56×56×64	56×56×128	3×3,1
Conv2b	56×56×128	56×56×128	3×3,1
Pool2	56×56×128	28×28×128	2×2,2
Inception	28×28×128	28×28×256	—
Pool3	28×28×256	14×14×256	2×2,2
Inception	14×14×256	14×14×512	—
Pool4	14×14×512	7×7×512	2×2,2
Avgpool	7×7×512	1×1×256	—
FC	1×1×256	1×1×136	—

面部关键点位置信息获取过程介绍如下：

1）输入人脸图像$\boldsymbol{I}$，经过第一级CNN获得偏移估计$\Delta \boldsymbol{S}_1$，加上初始关键点位置估计，得到该阶段的关键点位置估计$\boldsymbol{S}_1$；

2）计算$\boldsymbol{S}_0$相对于$\boldsymbol{S}_1$的仿射矩阵$\boldsymbol{T}_2$，$\boldsymbol{T}_2$用来将输入图像归一化到标准形状，得到矫正后的人脸图像$\boldsymbol{T}_2(\boldsymbol{I})$和关键点位置$\boldsymbol{T}_2(\boldsymbol{S}_1)$，并根据前一阶段产生的关键点位置估计值生成关键点热度图$\boldsymbol{H}_t$，本阶段$t=2$，计算公式如下：

$$\boldsymbol{H}(x,y)=\frac{1}{1+\min\limits_{s_i\in \boldsymbol{T}_t(\boldsymbol{S}_{t-1})}\left\|(x,y)-s_i\right\|} \tag{1}$$

其中：s_i为$\boldsymbol{T}_t(\boldsymbol{S}_{t-1})$第$i$个关键点位置，$(x,y)$为像素点位置，利用关键点热图，卷积神经网络可以推断出前一阶段估计的关键点位置。然后将$\boldsymbol{T}_2(\boldsymbol{I})$、$H_2$以及第一级全连接层的输出特征图$\boldsymbol{F}_2$作为第二级CNN的输入，得到$\Delta \boldsymbol{S}_2$，进一步利用$\Delta \boldsymbol{S}_2$和$\boldsymbol{T}_2(\boldsymbol{S}_1)$经过标志点变换获得该阶段的关键点位置估计$\boldsymbol{S}_2$，计算公式为：

$$\boldsymbol{S}_t=\boldsymbol{T}_t^{-1}\left(\boldsymbol{T}_t\left(\boldsymbol{S}_{t-1}\right)+\Delta \boldsymbol{S}_t\right) \tag{2}$$

3）计算误差：

$$E=\min_{\Delta \boldsymbol{S}_t}\frac{\left\|\boldsymbol{S}_t-\boldsymbol{S}^*\right\|}{d_{\text{ipd}}} \tag{3}$$

其中：$\boldsymbol{S}^*$为真实的关键点位置，d_{ipd}为$\boldsymbol{S}^*$的瞳孔距离。本文误差阈值设置为0.08，不满足则返回2）。

4）重复2）、3）阶段，直到训练完所有样本，保存训练好的面部关键点提取模型，批样本数为1 000，每批样本包含64张人脸图像。

5）利用面部关键点提取模型对真伪人脸进行定位，并将68个二维坐标转换为136维特征向量，输出面部关键点位置集合。

$$\begin{bmatrix}x_1,y_1\\x_2,y_2\\\vdots\\x_{68},y_{68}\end{bmatrix}\Rightarrow\begin{bmatrix}x_1 & y_1 & x_2 & y_2 & \cdots & x_{68} & y_{68}\end{bmatrix}^{\mathrm{T}} \tag{4}$$

在面部关键点的提取过程中，从初始关键点位置估计$\boldsymbol{S}_0$开始，生成的关键点热图$\boldsymbol{H}_t$、特征图$\boldsymbol{F}_t$和一个相似变换矩阵$\boldsymbol{T}_t$在网络的各个连续阶段之间形成链接。与之前的局部区域图像定位关键点的方法相比，增加了人脸的全局信息，避免被局部的信息误导，对有角度和部分遮挡状态下的人脸可以获得更为准确的定位，从而得到更好的检测效果。

1.3 分类网络

SVM是基于统计学学习理论提出的一种机器学习方法，主要思想是寻找一个超平面，使两组不同的高维数据尽可能地被超平面隔离，根据结构风险最小化原则，最大化分类间隔构建最优分类超平面来提高分类器的泛化能力，在统计样本量较少的情况下，也能获得良好效果。SVM与其他学习机相比，具有良好的泛化能力，在处理分类中的非线性、小样本、高维数等问题表现良好，适用于本文的分类工作。在应用SVM对真伪图像面部关键点进行分类前，先利用PCA对提取的真伪图像面部关键点位置数据进行降维处理，将136维的坐标数据转换为3维向量，以便利用SVM进行分类训练。PCA是一种使用广泛的数据分析方法，用于提取数据的主要特征分量。本文的SVM分类器应用LibSVM[15]开源模式识别软件包实现。SVM分类器的核函数和参数设置对分类效果有很大影响，尤其是核函数参数γ及分类惩罚因子C。本文采用LibSVM提供的径向基函数（Radial Basis Function，RBF）内核训练SVM分类系统：

$$\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\mathrm{e}^{-\gamma\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2};\ \gamma>0 \tag{5}$$

选用网格搜索法对参数进行调优，即对每个参数设置多个值，然后进行交叉验证尝试各种可能的(C,γ)组合，找出使交叉验证精确度最高的(C,γ)组合作为实际分类中的参数值。搜索范围设置为C:$[2^{-2},2^{-1},2^0,2^1,2^2]$，$\gamma$:$[10^{-4},10^{-3},$

$10^{-2}, 10^{-1}, 10^{0}$],通过五折交叉验证(5-fold cross validation),选择 $C = 2^{2}, \gamma = 10^{-2}$。同时考虑数据集中真伪人脸图像的数量分布不均衡,本文通过调整训练数据集中样本损失与类频率成反比的方法来平衡两个类的损失。

2　实验与结果分析

2.1　实验准备

实验配置如下:中央处理器(Central Processing Unit, CPU):Inter Core i7-9750H,主频为 2.60 GHz,内存为 16 GB;图像处理器(Graphic Processing Unit, GPU):GeForce RTX 2070MQ,显存为 8 GB。

实验中用到了三种数据集:300W、CelebA 和 PGGAN。300W 是一个标准人脸对齐数据集,由 Labeled Face Parts in the Wild (LFPW)[16]和 Annotated Faces Landmarks in the Wild (AFW)[17]等多个数据集综合而成,标注了 68 个关键点和瞳孔坐标,包含室内与室外两种情况下的人脸图像,整体难度系数较高。将 300W 作为训练集对基于 DAN 的面部关键点获取网络进行训练,获得的面部关键点提取模型具有较好的泛化能力。

CelebA 包含了 202 599 张的固定分辨率为 216×178 像素的真实人脸图像,广泛用于人脸相关的计算机视觉训练任务。

PGGAN 数据集由 10 000 张利用 PGGAN 技术伪造的 1 024×1 024 像素的人脸图像组成。

有角度及遮挡人脸数据集(以下简称非正脸数据集)来源于 CelebA 及 PGGAN 伪造的人脸图像。首先筛选出 CelebA 中有角度及遮挡人脸图像,然后对剩余图像随机选择 3 000 张进行手动遮挡,最后将这两个部分和 PGGAN 数据集进行合并,如图 4 所示;正脸数据集则由 PGGAN 和 CelebA 两个数据集直接合并而成,为了与非正脸数据集区分,将其简称为正脸数据集。

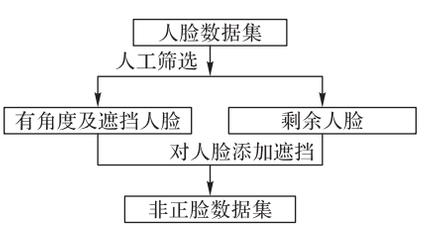


图 4　有角度及遮挡人脸数据集的获取

Fig. 4　Acquisition of faces with angles and occlusion

2.2　面部关键点提取网络准确度测试

为了验证本文面部关键点提取网络在有角度及遮挡情况下具有更高的定位准确度,使用 Dlib 函数库中的人脸检测器(以下简称为 Dlib)与本文网络分别提取有角度及遮挡人脸图像的 68 个面部关键点,并计算误差进行对比分析,其中人脸图像全部来自于用标准人脸数据集 300W。

将本文网络和 Dlib 提取的面部关键点位置与标准值求取误差并作曲线,结果如图 5 所示。可以看出,同样的数据,整体上本文网络的误差曲线要低于 Dlib,其中部分人脸关键点定位情况对比如图 6 所示。在有角度及遮挡的情况下,本文网络表现出更好的性能,而 Dlib 则容易被误导致使定位不准。定位不准意味着人为引入了面部不对称、不协调,从而影响对真伪人脸特别是真实人脸的检测效果。进一步根据如下公式计算均方根误差(Root Mean Square Error, RMSE):

$$R_{\mathrm{RMSE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{6}$$

其中:y_i 表示测试值,$\hat{y}_i$ 为真实值,共有 m 组数据。Dlib 的均方根误差为 77.78,远高于本文网络的 18.21,RMSE 越小,误差越低,效果越好,说明本文构建的人脸检测网络对有角度及遮挡人脸的关键点提取准确率更高。

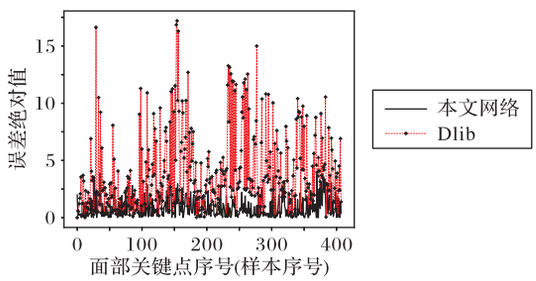
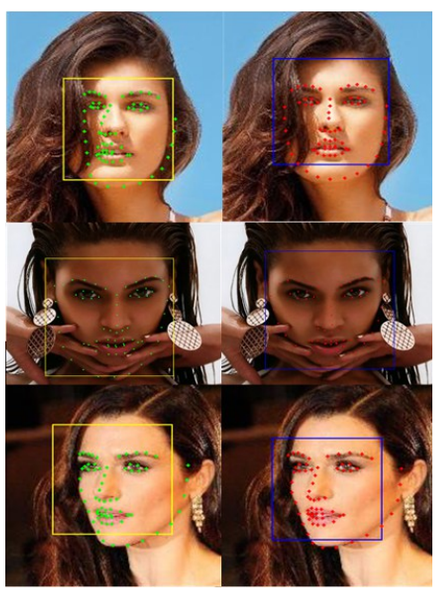


图 5　本文网络与 Dlib 误差曲线

Fig. 5　Error curves of proposed network and Dlib



(a) Dlib　　(b) 本文网络

图 6　Dlib 与本文网络对遮挡人脸的关键点定位情况

Fig. 6　Landmark location of faces with occlusion by Dlib and proposed network

2.3　GAN 伪造人脸检测方法实验评估

本节主要对本文提出的伪造人脸图像检测方法进行整体实验评估,为了避免阈值的选取影响分类结果,实验选择受试者工作特征(Receiver Operating Characteristic, ROC)曲线下面积(Area Under ROC, AUC)和精确率召回率(Precision Recall, PR)曲线下面积(Average Precision, AP)为性能指标。ROC 的 X 轴为伪阳性率(False Positive Rate, FPR),表示 SVM 分类器判断为真实人脸中实际为 GAN 伪造人脸占所有 GAN 伪造人脸的比例;Y 轴为真阳性率(True Positive Rate, TPR),表示分类模型判断为真实人脸中实际为真实人脸占所有真实人脸的比例。PR 曲线的 Y 轴表示精确率(Precision),X 轴表示召回率(Recall)。

首先利用 300W 数据集对面部关键点信息提取网络进行训练,保存模型;接着使用训练好的模型提取正脸数据集和有角度及遮挡人脸数据集的面部关键点。为了对比 PCA 降维处理对实验结果的影响,对关键点数据分别做以下两种处理:1) 利用 PCA 对 136 维的坐标数据进行降维处理转换为三维特征向量,以便进行分类;2)将每组数据展开为一维数组。然后将处理后的正脸数据集输入 SVM 分类器进行训练,输出训

练结果，并保存分类模型；最后输入非正脸数据集得到测试结果。图 7 和图 8 分别显示了本方法与其他基于深度神经网络的检测方法在正脸数据集和非正脸数据集上的 ROC 曲线和 PR 曲线，量化结果如表 2 所示，包括在相同数据集上使用不同的分类方法和人脸检测网络的综合性能对比。

实验结果表明，在正脸和非正脸情况下，本文方法对真伪人脸的分类效果均优于其他检测方法，包括一些基于深度神经网络的方法（如 VGG19 和 XceptionNet），并且这两种基于深度神经网络的方法还需要较大的参数量和更高的硬件要求。在测试集为有角度和遮挡人脸的情况下，各检测方法的 AUC 值和 AP 都出现了不同程度的下降，但本文方法的 AUC 值还是达到了 89.76%，AP 也保持在 96.38%，明显优于其他检测方法。虽然由于面部关键点提取网络的参数个数较多导致本文方法参数量大于文献[9]方法的参数量，但还是远小于另外两种基于深度神经网络方法的参数量；并且面部关键点提取网络可以单独进行训练，不影响检测效果。当分类方法相同时，未经 PCA 处理的 AUC 值均略有下降，并且同样对 302 599 张真伪人脸进行分类，未经 PCA 处理时运行耗时增加 5 945 ms，存储数据也需要更大空间，因此在图像较多时选择 PCA 对数据进行处理更优。同样进行 PCA 处理，DAN 进行改进后的 AUC 值分别提升了 1.91% 和 2.92%，并且有更小的参数量和计算量，性能优于原网络。

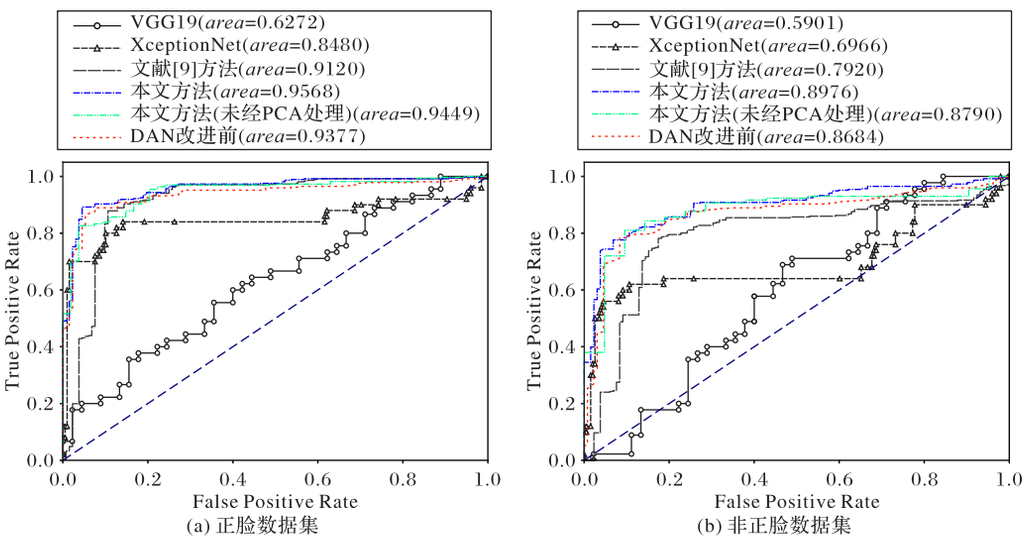


图 7 不同检测方法的 ROC 曲线

Fig. 7 ROC curves of different detection methods

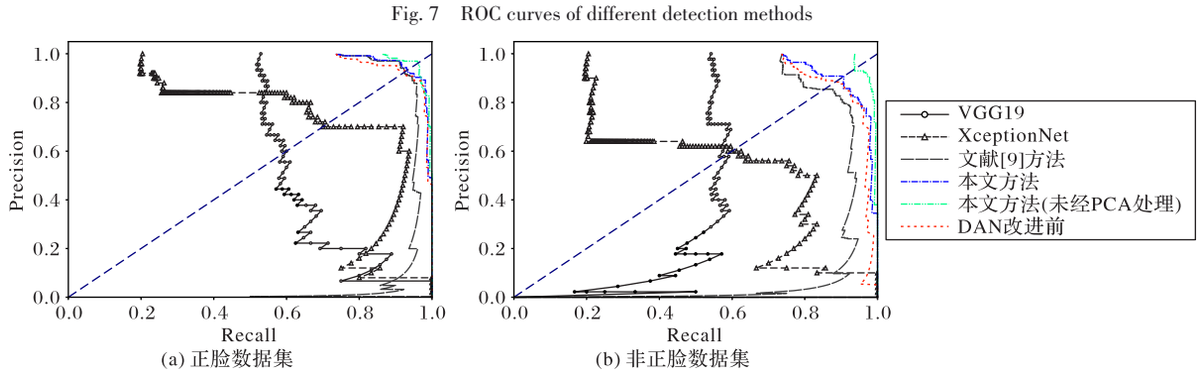


图 8 不同检测方法的 PR 曲线

Fig. 8 PR curves of different detection methods

表 2 不同检测方法的综合性能对比

Tab. 2 Overall performance comparison of different detection methods

方法	参数量/10^6	AUC/%		AP/%		是否需要 GPU
		正脸数据集	非正脸数据集	正脸数据集	非正脸数据集	
VGG19	~143.7	62.72	59.01	65.32	53.93	是
XceptionNet	~22.9	84.80	69.66	76.03	58.76	是
文献[9]方法	~0.11[9]	91.20	79.20	94.18	88.96	否
本文方法	~4.2	95.68	89.76	98.44	96.38	否

综上，本文提出的 GAN 伪造人脸图像检测方法有三个优点：1）面部关键点位置与图像大小无关，在训练和使用所获得的分类方法时不需要重新对图像进行缩放，可以有效避免由于缩放操作而导致的影响；2）基于该特征向量的分类方法对硬件需求不高，参数量小，计算成本较低；3）该方法旨在提升检测正确率，在正脸和有角度和遮挡人脸图像上都表现出了更好的性能，并不只适用于某一类人脸图像。

3 结语

针对现有 GAN 伪造图像检测方法对于有角度及遮挡情

况下的人脸考虑较少,存在真实人脸误判问题,本文提出基于DAN来提取68个面部关键点,获得更为准确的面部位置信息,进而通过SVM对面部关键点特征进行分类甄别真伪人脸图像,降低了有角度及遮挡人脸的误判率。最后实验验证了本文方法相比其他检测方法在正脸和有角度及遮挡人脸上都展现出了更好的性能,能有效提高GAN伪造人脸图像的检测正确率。由于GAN模型缺乏对不同人脸部件配置的约束这一问题目前还未得到彻底修正,因此,基于面部关键点的方法还可以进行进一步研究。例如,在真实人脸面部表情强烈扭曲的情况下如何避免误判,以及进一步提升对伪造人脸图像的甄别效果,下一步可以考虑通过获取面部的三维空间位置来进行分类,暴露三维空间中GAN伪造人脸的缺陷来提高检测正确率。

参考文献(References)

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[2] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation [EB/OL]. [2020-11-11]. https://arxiv. org/pdf/1710. 10196. pdf.

[3] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4396-4405.

[4] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN [C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 8107-8116.

[5] LI H, LI B, TAN S, et al. Identification of deep network generated images using disparities in color components [J]. Journal of Signal Processing, 2020, 174: No. 107616.

[6] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues [C]// Proceedings of the 2019 IEEE International Conference on Image Processing. Piscataway: IEEE, 2019: 4584-4588.

[7] NATARAJ L, MOHAMMED T M, MANJUNATH B S, et al. Detecting GAN generated fake images using co-occurrence matrices [J]. Electronic Imaging, 2019, 2019(12): No. 532.

[8] TARIQ S, LEE S, KIM H, et al. Detecting both machine and human created fake face images in the wild [C]// Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. New York: ACM, 2018:81-87.

[9] YANG X, LI Y, QI H, et al. Exposing GAN-synthesized faces using landmark locations [C]// Proceedings of the 2019 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 113-118.

[10] KOWALSKI M, NARUNIEC J, TRZCIŃSKI T. Deep alignment network: a convolutional neural network for robust face alignment [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2034-2043.

[11] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 3730-3738.

[12] KAZEMI, VSULLIVAN J. One millisecond face alignment with an ensemble of regression trees [C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 1867-1874.

[13] LEVER J, KRZYWINSKI M, ALTMAN N. Principal component analysis[J]. Nature Methods, 2017, 14(7): 641-642.

[14] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.

[15] CHIH C C, CHIH J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): No. 27.

[16] BELHUMEUR P N, JACOBS D W, KRIEGMAN D J, et al. Localizing parts of faces using a consensus of exemplars [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2930-2940.

[17] ZHU X, RAMANAN D. Face detection, pose estimation, and landmark localization in the wild [C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 2879-2886.

[18] YNAG X, LI Y, LYU S. Exposing deep fakes using inconsistent head poses [C]// Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2019: 8261-8265.

[19] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2020-11-11]. https://arxiv. org/pdf/1511. 06434. pdf.

[20] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2017: 5769-5779.

[21] ARJOVSKY M, CHITALA S, BOTTOU L, et al. Wasserstein generative adversarial networks [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 214-223.

[22] SUN Y, WANG X, TANG X. Deep convolutional network cascade for facial point detection [C]// Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 3476-3483.

[23] 颜普,苏亮亮,邵慧,等. 基于多支持区域局部亮度序的图像伪造检测[J]. 计算机应用, 2019, 39(9):2707-2711. (YAN P, SU L L, SHAO H, et al. Image forgery detection based on local intensity order and multi-support region [J]. Journal of Computer Applications, 2019, 39(9): 2707-2711)

This work is partially supported by the National Key Research and Development Program of China (2016YFB0501900).

TANG Guihua, born in 1996, M. S. candidate. Her research interests include image processing, computer vision.

SUN Lei, born in 1973, Ph. D., professor. His research interests include cyberspace security, trusted computing.

MAO Xiuqing, born in 1980, M. S., associate professor. His research interests include information security.

DAI Leyu, born in 1990, M. S., lecturer. His research interests include neural network encryption, hardware acceleration.

HU Yongjin, born in 1981, M. S., lecturer. His research interests include active defense, situation awareness.